ML Challenge 2025

Smart Product Pricing Solution

Team Name:	SheCodesML
Team Members:	Jiya Sinha, Himadri Sharma, Rashi Bharti
Date of Submission:	13/10/2025

1 Summary

We found that textual descriptions of a product are more informative than visual features. We leveraged **Sigmoid Loss for Language-Image Pre-training (SigLIP)**, a pretrained multimodal image–text representation model, and fine-tuned its **text projection head and final transformer layer** to adapt textual embeddings for the pricing domain. The image embeddings were kept frozen to preserve visual generalization. A lightweight **Multi-Layer Perceptron (MLP)** was trained on the fused image–text representations to predict log-transformed prices efficiently. This setup allowed effective multimodal learning while minimizing computational cost and overfitting.

2 Methodology Overview

2.1 Problem Analysis

We analyzed the pricing dataset to identify patterns and challenges in combining textual and visual data.

2.1.1 Key Observations

- The catalog entries showed high variability, some product descriptions were detailed while others were minimal or redundant.
- Product images varied in size and composition. Most lacked informative text or visual cues, making packaging appearance the primary visual signal.
- Textual descriptions, on the other hand, carried rich semantic information about product type, brand, and quantity making them more predictive of price.
- There were relatively few high-priced products than low-priced ones.

2.2 Solution Strategy

We approached the problem as a multimodal regression task, integrating both text and image modalities to predict product prices, with higher weightage given to textual information. To handle the skewed price distribution, since there were relatively few high-priced products, we applied a logarithmic transformation for stable regression. Model performance was evaluated using SMAPE on the original (exponentiated) price scale.

Approach Type: Hybrid (Multimodal — Text + Image using MLP)

Core Innovation: Multimodal image-text model, SigLIP text fine-tuning, fused image–text representations.

3 Model Architecture

3.1 Architecture Overview

Refer to the figure 1 for our model architecture.

3.2 Model Components

- **Image Encoder**: Frozen SigLIP vision layer.
- **Text Encoder**: SigLIP text tower with last transformer layer and projection head fine-tuned.
- **Fusion Layer**: Weighted average of normalized image and text embeddings.
- **Regression Head**: 3-layer MLP with GELU activations and dropout for regularization

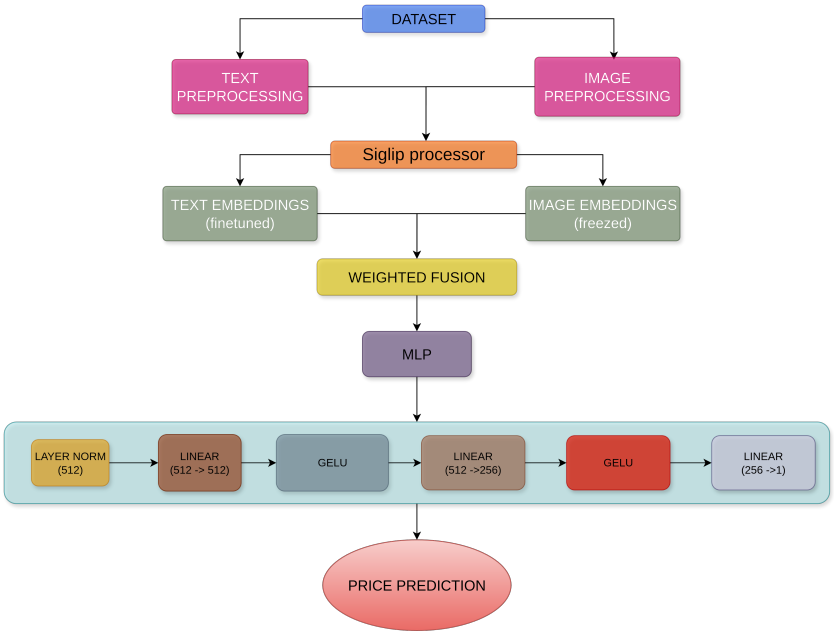

Figure 1: **Proposed Architecture:** The model consists of a pretrained **SigLIP encoder** for image and text embeddings. Both embeddings are L2-normalized and fused using a **weighted combination** (0.3×image + 0.7×text). The fused embedding is passed through a 3-layer MLP regression head that predicts log-transformed prices.

3.2.1 Text Processing Pipeline:

Steps

- Convert all text to lowercase, remove URL, replace multiple dots with a single dot.
- Normalize spaces (convert extra spaces to a single space) and remove non-alphanumeric characters.
- Extract structured information (Item Name, Description, Bullet Points, Value, Quantity) from the catalog content.
- Merge these extracted fields into a single sentence representing each product.
- Remove repeated phrases to reduce redundancy.

3.2.2 Image Processing Pipeline:

- Download product images from URLs in the dataset and resize the images uniformly to 128x128 pixels.
- Save each image in .jpeg format using a unique hashed filename and maintain consistent quality by saving at 80% JPEG quality for optimized storage balance.

4 Model Performance

The model was trained using **L1 Loss** on the log-transformed price and **evaluated using SMAPE** on the original price scale. The best validation SMAPE achieved was **47%**, demonstrating robust alignment between predicted and true prices.

5 Conclusion

By fine-tuning text embeddings and training a lightweight MLP on fused image–text features, we achieved stable and interpretable performance with a best SMAPE of 47%. The key takeaway is that textual information contributes more strongly to pricing accuracy than visual cues, highlighting the importance of structured language understanding in multi-modal learning.